LOD 的网络结构分析与可视化*

夏立新　谭　荧
(华中师范大学信息管理学院　武汉　430079)

摘要:【目的】对关联开放数据(LOD)进行结构特征分析，利用分析结果指导关联数据的组织实践。【方法】通过度分布、平均路径长度、聚类系数等指标描述 LOD 网络结构，对比复杂网络理论中的两个基本性质：无标度特性和小世界效应。【结果】LOD 整体网络结构具有近似无标度网络的幂率分布特征，图书馆学、情报学领域子网具有相对均匀的指数分布特征，两网同时具有短平均路径长度和高聚类系数的小世界效应。【局限】缺乏对关键节点的多权重赋值。【结论】LOD 的小世界特性能优化检索效率，而无标度特性会降低整个网络的稳定性。

关键词: LOD　复杂网络　网络结构　可视化
分类号: G203

1　引　言

越来越多的数据拥有者将他们的数据以关联数据的形式发布到网络上，形成一个全球化的数据空间，即数据网络(Web of Data)[1]。相对于传统的文档网络，数据网络更加结构化，将简单的超链接变成了复杂的关系网，使 Web 上的数据能够被发现和检索，并被人和机器所理解。2014 年 8 月，W3C 关联开放数据项目发布了最新的关联开放数据云图(Linked Open Data Cloud, LOD Cloud)，为数据网络建立了一个视觉模型，该图绘制的开放关联数据集由最初的几十个增长到几百个，内容涵盖媒体、政府、出版物、地理、生命科学、跨领域、用户生成内容和社交网络 8 个领域[2]。LOD 云图将不同领域的关联开放数据资源整合为一个互联的网络并将其可视化，从情报学的视角来看，这是继引证、共词、合著等典型知识网络之后的一种新型网络型态。那么，LOD 网络具有怎样的结构属性？各数据集之间的连接是否有特殊规律与特征？对于此类问题的研究，有助于认识和评价关联数据的发展现状，指导实践中关联数据的发布、互联和检索。

目前国内对关联数据的研究主要集中在发布技术[3-6]、互联方法[7-9]和资源集成[10-11]等方面，尚无以数据集为基本单元对整个关联数据网络结构进行研究。国外已有一些相关研究，Schmachtenberg 等统计了这些年关联开放数据集的增长和互联，认为关联开放数据网络已经由以 DBpedia 为核心的结构转化为更加分散的非中心性结构，关联数据在数量以几何级增长的同时，内容也逐步多元化发展[12]。Auer 等通过统计关联数据集有多少有效的出入链接来评价数据集的质量，统计过程中经常遇到运行中断、限制获取和非标准 SPARQL 终端等问题，因此他们认为现有关联数据统计数字过于乐观，网上实际可用的关联数据集比其统计数据要低一个数量级[13]。Campinas 等基于语义网搜索引擎 Sindice，对关联数据集中本体、谓词、字符串和 URI 等数据进行统计，为评价实体导向的语义搜索系统提供数据支持[14]。Bizer 等通过对微数据、微格式和 RDFa 三种标记方法的利用率进行比较分析，展示网页中结构化数据的分布和发展[15]。

上述文献从不同角度对关联数据集进行统计和分析，一定程度上描述了关联数据网络的发展现状。关联数据的 RDF (Resource Description Framework)数据模型，使其拥有典型的网络拓扑结构特征。本文利用复杂网络理论中度分布、平均路径长度、聚类系数等拓扑性质描述关联开放数据的结构，从网络联系的角

通讯作者：谭荧，ORCID：0000-0002-7987-4696，E-mail：735014860@qq.com。
*本文系国家社会科学基金重大招标项目“基于多维度聚合的网络资源知识发现研究”(项目编号:13&ZD183)的研究成果之一。

度分析如何有效组织关联数据，揭示隐含在结构表象下的潜在关系。

2 复杂网络基本性质

(1) 无标度特性

网络的度分布是刻画无标度网络的重要参数。ER随机网络的度分布近似为泊松分布，而大部分复杂网络度分布具有幂率形式。一般认为度分布服从幂率分布并且幂指数 r≤3 的网络是无标度网络，这种性质也称为无标度特性。网络的无标度特性一般按照其度、出度和入度分别考虑。人们对不同领域的现实网络研究发现，包括万维网和引文网络在内的许多网络节点度分布都满足幂率分布[16]。

(2) 小世界网络

网络的平均路径长度，是指所有可达节点对之间最短路径的加权平均值；网络的平均聚类系数是指所有节点聚类系数的平均值。与具有相同节点数和平均度的随机网络相比，既具有较短的平均路径长度，又具有较高的聚类系数的网络称之为小世界网络[17]。

3 数据和方法

3.1 数据采集

本文的数据来自 Datahub[18]，一个基于数据管理系统 CKAN 的免费数据管理平台。Datahub 对其中近 9 千多个数据集进行分组和标记标签。其中，以关联开放数据形式发布，并且与其他数据集有连接的数据集分组为 LOD Cloud Group，该组也是 LOD 云图中数据集的主要来源。将 LOD Cloud 整组数据集作为研究对象，利用 datahub2void 软件获取数据集的 VoID 词表[19]，提取日期截止到 2015 年 4 月 28 日。

在关联数据中，不仅每个数据集内部存在着链接，数据集之间也存在着联系。一个数据集发布后，应该确保有外部的RDF链接指向这个数据集的URIs，这样新数据集才能被RDF浏览器和爬虫发现，才能补充现有数据集的资源[20]。本文用一个节点代表一个数据集，节点之间的连线即数据集之间的链接，RDF链接是有向的，故关联开放数据网络是一个有向网络。本文将LOD Cloud组的数据集构成的网络称为LOD Cloud，其中带有Publication标签的数据集是由图书馆数据集、科学出版物、会议、大学读物、引文数据集组成，笔者将其称为图书馆学、情报学领域的关联开放数据，提取为Publication网络进行对比分析。Publication子网络的连线定义为子网内部节点之间的连线。具体统计数据如表1所示：

表 1 LOD Cloud 和 Publication 网络

网络	节点(N)	有向线(M)	平均度(k)
LOD Cloud	258	958	3.71
Publication	79	527	6.67

3.2 分析方法

(1) 指标计算

①累积度分布

有向网络的入度分布 $P(k^{in})$是网络中随机选取的一个节点的入度为 k^{in} 的概率。出度分布 $P(k^{out})$为网络中随机选取的一个节点的出度为 k^{out} 的概率。度分布 P(k)表示随机选取的节点度为 k 的概率。为了更清晰展示度分布图，本文使用累积度分布P_k来绘制度分布图，它表示的是度不小于k的节点的概率分布。

$$P_k = \sum_{k'=k}^{\infty} P(k') \tag{1}$$

如果度分布为幂率分布，即 $P(k) \propto k^{-\gamma}$，那么累积度分布函数 P_k 符合幂指数为γ–1 的幂律：

$$P_k \propto \sum_{k'=k}^{\infty} k'^{-\gamma} \propto k'^{-(\gamma-1)} \tag{2}$$

如果度分布为指数分布，$P(k) \propto e^{-k/\kappa}$，其中 κ >0，那么累积度分布函数 P_k 也是指数型的，且具有相同的指数：

$$P_k \propto \sum_{k'=k}^{\infty} e^{-k'/\kappa} \propto e^{-k/\kappa} \tag{3}$$

②平均路径长度

网络中任意两个节点 i 和 j 之间的距离 d_{ij}，定义为连接这两个节点的最短路径上的边数。网络的平均路径长度 L 定义为任意两个节点之间的距离的平均值，即：

$$L = \frac{1}{\frac{1}{2}N(N+1)} \sum_{i \geqslant j} d_{ij} \tag{4}$$

网络平均路径长度的概念严格来说只对连通图才是有限值，但很多现实网络都是不连通的，本文采用有向网络的经典算法，Dijkstra 算法。对于同等规模的随机图网络，设网络的节点数为 N、平均度为 k，那么该网络的平均路径长度为：

$$L_{random} \approx \ln N / \ln k \tag{5}$$

③平均聚类系数

假设网络中的一个节点 i 有 k_i 条边将它与其他节点相连，这 k_i 个节点就称为节点 i 的邻居，邻居节点之间实际存在的边数 E_i 和这 k_i 个节点之间最多可能的边数 $k_i(k_i-1)/2$ 之比就定义为节点 i 的聚类系数 C_i。整个网络的聚类系数 C 就

是所有节点 i 聚类系数 C_i 的平均值，即：

$$C = \frac{1}{N}\sum_{i=1}^{N}\frac{2E_i}{k_i(k_i-1)} \tag{6}$$

对于节点数为 N、平均度为 k 的随机图网络，平均聚类系数为：

$$C_{random} \approx k/N \tag{7}$$

(2) 相关分析

本文研究的变量为定序变量，通过计算斯皮尔曼(Spearman)等级相关性系数分析变量之间的相关性，相关性系数ρ在 0.00±0.30 为微相关，在±0.30~±0.50 为实相关，在±0.50~±0.80 之间为显著相关，在±0.80~±1.00 之间为高度相关，显著性水平$\rho<0.05$具有统计学意义。

(3) 回归分析

为判定指标的分布形态，在 Matlab 中绘制上述数据的散点图，利用 Curve Fitting Tool 添加拟合曲线。根据 SSE(误差平方和，趋向 0 最好)、R-Square(确定系数，趋向 1 最好)、Adjusted R-Square(调整确定系数，趋向 1 最好)和 RMSE(标准差，越向 0 最好)选择最佳拟合函数，依据拟合函数判定指标分布规律。

(4) 可视化

利用 Gephi 绘制 LOD Cloud 和 Publication 网络的结构图，利用不同颜色的节点代表不同领域的数据集，节点的大小代表度的大小，有向连线代表数据集之间的连接，连线的粗细代表连接的权重。

4　结果分析

4.1　度与相关性

(1) 入度与出度

本文采集的数据显示，LOD Cloud 网络 89%的节点度数不为零，Publication 子网中也有 77%的节点度数不为零，这一方面说明关联开放数据集并不孤单，另一方面也说明关联数据集之间的连接还有很大的发展空间。表 2 和表 3 列出了两个网络中出度和入度前 10 的数据集。

表 2　LOD Cloud 入度和出度前 10 的数据集

排名	数据集	入度	排名	数据集	出度
1	DBpedia	140	1	DBLP (RKBExplorer)	35
2	GeoNames	37	2	ePrints (RKBExplorer)	31
3	ePrints (RKBExplorer)	27	3	ACM (RKBExplorer)	31
4	DBLP (RKBExplorer)	27	4	ECS Southampton (RKBExplorer)	31
5	ACM (RKBExplorer)	26	5	DBpedia	29
6	Freebase	24	6	Wiki (RKBExplorer)	29
7	CiteSeer (RKBExplorer)	24	7	CiteSeer (RKBExplorer)	27
8	Wiki (RKBExplorer)	24	8	RAE2001 (RKBExplorer)	27
9	ECS Southampton (RKBExplorer)	24	9	KISTI (RKBExplorer)	25
10	OAI (RKBExplorer)	23	10	Newcastle (RKBExplorer)	25

表 3　Publication 入度和出度前 10 的数据集

排名	数据集	入度	排名	数据集	出度
1	ePrints (RKBExplorer)	26	1	ePrints (RKBExplorer)	30
2	ACM (RKBExplorer)	25	2	DBLP (RKBExplorer)	30
3	DBLP (RKBExplorer)	25	3	ACM (RKBExplorer)	28
4	OAI (RKBExplorer)	23	4	ECS Southampton (RKBExplorer)	27
5	CiteSeer (RKBExplorer)	23	5	CiteSeer (RKBExplorer)	26
6	Wiki (RKBExplorer)	23	6	Wiki (RKBExplorer)	26
7	RAE2001 (RKBExplorer)	22	7	RAE2001 (RKBExplorer)	25
8	ECS Southampton (RKBExplorer)	22	8	KISTI (RKBExplorer)	24
9	dotAC (RKBExplorer)	21	9	Newcastle (RKBExplorer)	24
10	KISTI (RKBExplorer)	21	10	LAAS (RKBExplorer)	22

其中DBpedia以140的入度排名第一，也就是说它被LOD网络中大部分的数据集指向，说明它具有丰富的数据资源并且涉及领域广泛，对于后发布的数据集是一个可信任的链接资源。GeoNames作为全球地理数据库同样具有很高的入度。这种节点更倾向与那些拥有较高连接度的“大”节点相连的现象，表明关联开放数据网络具有“优先连接”特性。然而DBpedia和GeoNames的出度相比入度而言则小很多，这与它们的发布时间较早有关。同时反映了关联数据网络存在的一个普遍问题，很多关联数据集在发布之后缺少维护，没有及时链接新发布的数据集，失效的链接也没有及时修订，从而降低了整个LOD网络的连通性。

表3中近一半的数据集也出现在表2的排名中，意味着Publication中高度数的节点相比网络其他领域的核心节点，度数也较高。然而这些节点的度数在表3中与表2中相差并不大，也就是说Publication中高度数的节点连接更倾向于连接领域内节点，在连接整个网络其他节点上的贡献并不大。

(2) 入度与出度相关性

整个LOD网络节点的入度和出度的Spearman相关性系数ρ=0.6546，显著性水平$\rho=5.98\times10^{-33}<0.05$，即关联数据集的出度和入度显著相关。表3中的排名显示Publication网络中入度和出度排名较高的节点很多是一样的，即核心节点同时具有较高的入度和出度。

Publication网络节点入度和出度的Spearman相关性系数ρ=0.8939，显著性$\rho=1.5\times10^{-28}<0.05$，即图情领域数据集的出度和入度高度相关。入度与出度的正相关说明关联数据集倾向于连接其他数据集更常连接的数据集。

4.2 累计度分布

从图1的拟合情况看，LOD Cloud的累积入度、累积出度和累积度分布都近似幂率分布且幂指数$r\leqslant3$，可以认为LOD Cloud网络具有无标度网络特性。大部分的节点(28%)度数为1，分布尾部稀少，即存在少量节点被大多数节点连接。具有这样结构特征的网络，即使局部节点失效，也不会影响整个网络的稳定性，但高度数节点失效，会导致整个网络非常脆弱，信息不能顺畅流通。也就是说LOD Cloud网络中少数最受欢迎的节点起到了连接大部分节点的重要作用，找到这些节点与之关联能快速加入关联数据网络的最大连通片，共享更多资源。然而如果新节点都倾向与高度数的中心节点连接，一旦中心节点失效，可破坏整个网络的连通性。

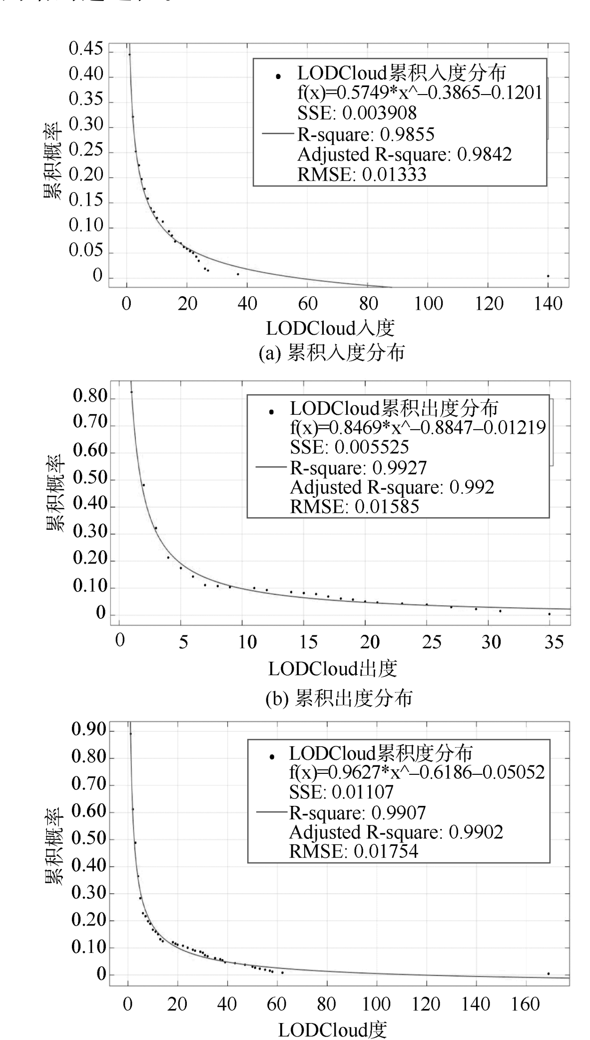


图1 LOD Cloud网络的累积度分布与拟合曲线

图2显示图书馆学、情报学领域子网的累积入度、累积出度和累积度分布都近似指数分布。随着度的增加，累积概率并没有急剧减少或增加，意味着Publication网络的度分布相对均匀。这样的网络结构具有更强的稳定性，网络连通并不依赖于少数度数极高的节点，即使局部节点失效，对整个网络连通性影响也不大。Publication度分布并未继承LOD Cloud的无标度特性，说明关联开放数据各领域的网络结构存在差异性，并不是简单的叠加。

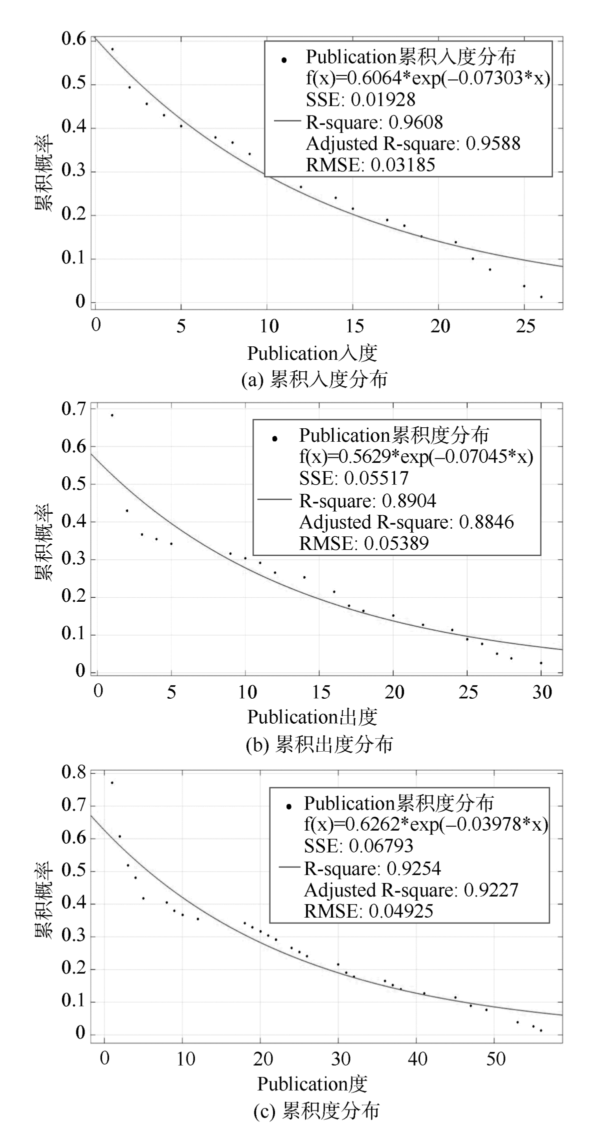


图 2　Publication 累积入度、累积出度及累积度分布

4.3　平均路径长度和聚类系数

由表4可知，两个网络的平均路径长度L小于同等规模的随机网络的平均路径长度L_{random}，但是聚类系数C远大于C_{random}，也就是说LOD Cloud和Publication具有明显的小世界网络特征。短平均路径长度意味着即使关联开放数据集不断增加，数据集之间的距离却很近，这样的结构能使检索时间加快。高聚类系数表明关联数据的连接并不随机，与数据集A相连的两个数据集B和C，彼此也相连的可能性很高。这样的结构使描述同一实体的资源互相连接，丰富了描述实体的多样性。简言之，小世界特性使网络既能保证快速找到数据，又能保证数据的丰富性，关联开放数据的结构有利于提高检索效率。

表 4　平均路径长度和聚类系数

网络	L	C	L_{random}	C_{random}
LOD Cloud	2.40	0.2391	4.23	0.0143
Publication	1.51	0.3138	2.30	0.0844

5　可视化分析

由图 3 可以清晰看到 Publication 领域数据集紧密聚集在一起，此外生命科学领域数据集也组成一个小型的连通片，其他领域数据集则没有形成明显的聚集，说明关联数据集在科学领域互联的应用较多。图 3 左上部媒体领域有两个以英国 BBC Music 和 Music Brainz 为中心节点的星型拓扑，但它们都未和美国著名媒体 New York Times 连接，三者之间的最短路径也是通过 DBpedia 实现，意味着同一领域业界巨头发布的关联数据彼此互不相连的原因可能是地域的分隔。政府领域关联数据集也印证了这点，除了英国政府发布的几个数据集互相有连接之外，其他政府数据集都各自孤立。用户生成内容和社交网络领域内的数据集也是完全分散，彼此毫无联系。跨领域的数据集连接较为多样，最常见的是与地理的数据集相连。这种复杂的网络连接意味着关联数据并不能按照数据集的领域划分层次，笔者认为要使关联数据更为紧密，需要连接的是发布关联数据集的机构和人。

有研究表明，数据集之间最常用的连接谓词是 owl:sameAs 和 rdfs:seeAlso，用于连接描述同一对象的两个资源[12]。图书馆学、情报学领域的数据多为书目、论文、作者和研究机构，这些信息在各数据集中有很多重复，故容易形成较多互联，形成紧密关联。图 4 展示了 Publication 中高度数的节点更倾向与高度数节点互联。其中的强连通片是由利用 RKB Explorer 发布的数据集组成，甚至美国国会图书馆发布的 LCSH 数据集也没有形成这样大的连通片。RKB Explorer 应用的底层架构使用一致引用服务(Consistent Reference Services, CRS)实现指向同一事物的 URIs 的连接[21]，由此推断关联开放数据的互联还存在技术上的阻隔[7]。

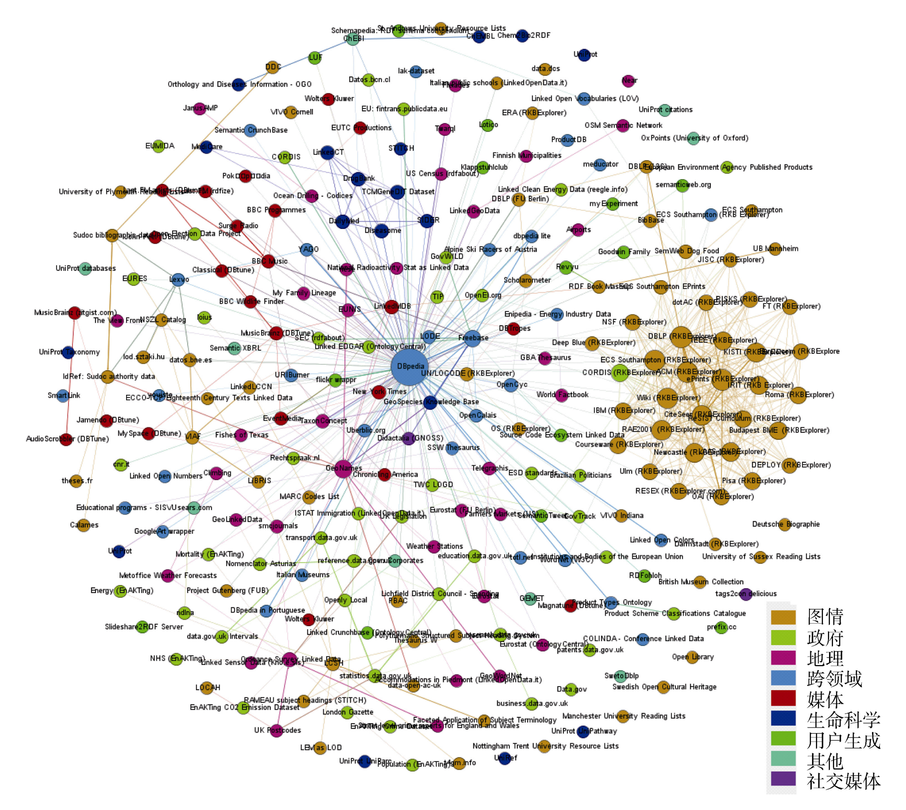


图 3　LOD Cloud 网络结构图

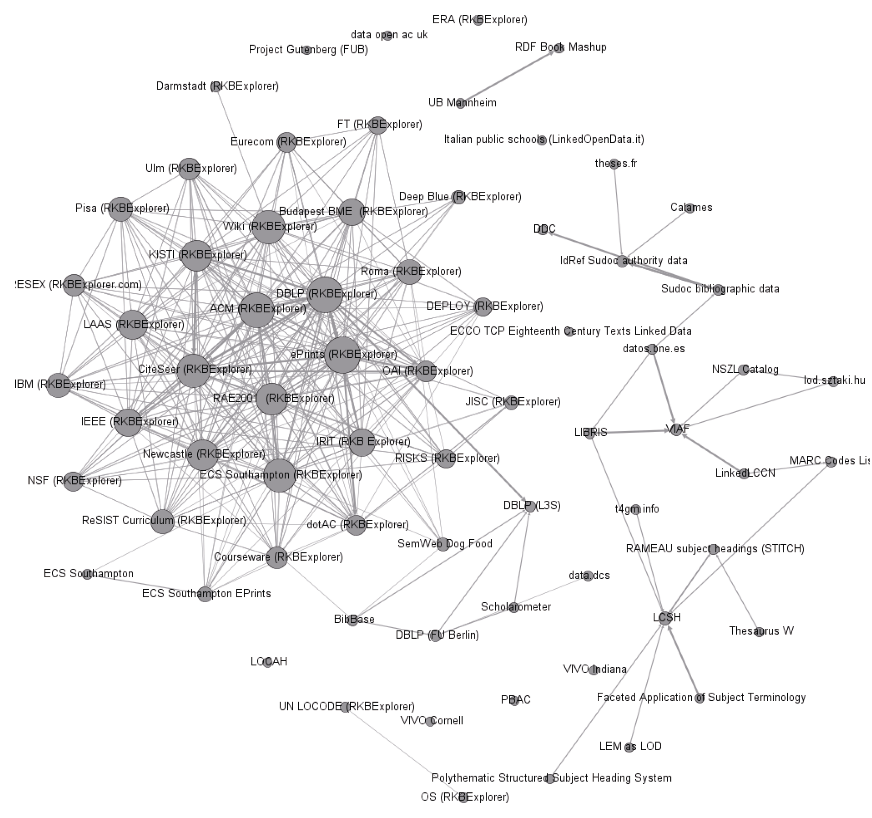


图 4　Publication 网络结构图

6　结　语

关联开放数据网络结构在整体层面上具有近似无标度网络的幂率分布特征，同时具有短平均路径长度和高平均聚类系数的小世界特性。图书馆学、情报学领域的关联数据网络具有相对均匀的指数分布特征，同时具有小世界网络特性。小世界网络的共性能帮助关联开放数据优化检索效率，然而倾向连接高度数节点的趋势会使整个关联数据网络的稳定性降低，故发布数据集时要慎重选择数据集互联。关联数据网络结构图显示层级结构与领域内容并无关联，地域和技术的不同是关联数据网络连接不紧密的重要因素。

未来关联开放数据的网络结构研究可以进行以下工作：权重是非常重要的统计指标，对关键节点权重赋值有助于更深一步了解关联数据网络的结构特性；目前对网络结构的研究还停留在静态的统计分析上，信息的结构会随着时间而改变，新的数据集会产生新的属性，关联开放数据网络也在演化，对演化过程的研究会帮助人们更全面地认识关联开放数据。

参考文献：

[1] Bizer C, Heath T, Berner-Lee T. Linked Data-The Story So Far [J]. International Journal on Semantic Web and Information Systems, 2009, 5(3): 1-22.

[2] Schmachtenberg M, Bizer C, Paulheim H. State of the LOD Cloud 2014 [R/OL]. (2014-08-30). [2015-04-28]. http:/linked-datacatalog.dws.informatik.uni-mannheim.de/state/.

[3] 夏翠娟，刘炜，赵亮，等. 关联数据发布技术及其实现——以 Drupal 为例[J]. 中国图书馆学报，2012，38(1)：49-57. (Xia Cuijuan, Liu Wei, Zhao Liang, et al. The Current Technologies and Tools for Linked Data: A Case of Drupal[J]. Journal of Library Science in China, 2012, 38(1): 49-57.)

[4] 沈志宏，刘筱敏，郭学兵,等. 关联数据发布流程与关键问题研究——以科技文献、科学数据发布为例[J]. 中国图书馆学报，2013，39(2)：53-62. (Shen Zhihong, Liu Xiaomin, Guo Xuebing, et al. A Research on Publishing Workflow and Key Issues of Linked Data: Experience with Publishing Scientific Literature and Scientific Data as Linked Data [J]. Journal of Library Science in China, 2013, 39(2): 53-62.)

[5] 王忠义，夏立新，石义金，等. 数字图书馆中层关联数据的创建与发布[J]. 现代图书情报技术，2013(5)：28-33. (Wang Zhongyi, Xia Lixin, Shi Yijin, et al. The Creation and Publishing of Middle Linked Data in Digital Library [J]. New Technology of Library and Information Service, 2013(5): 28-33.)

[6] 白海燕，梁冰. 利用 D2R 实现关系数据库与关联数据的语义模式映射[J]. 现代图书情报技术，2011(7-8)：1-7. (Bai Haiyan, Liang Bing. Semantic Pattern Mapping Between RDBMS and Linked Data Based on Open Source Software [J]. New Technology of Library and Information Service, 2011(7-8): 1-7.)

[7] 沈志宏，黎建辉，张晓林. 关联数据互联技术研究综述：应用、方法与框架[J]. 图书情报工作，2013，57(14)：125-133. (Shen Zhihong, Li Jianhui, Zhang Xiaolin. Research Review on the Interlinking Technology of Linked Data: Applications, Methods and Frameworks [J]. Library and Information Service, 2013, 57(14): 125-133.)

[8] 朱雯晶，夏翠娟，刘炜. SILK 关联发现框架综析[J]. 现代图书情报技术，2013(4)：18-24.(Zhu Wenjing, Xia Cuijuan, Liu Wei. Analysis of SILK Linkage Discovery Framework[J]. New Technology of Library and Information Service, 2013(4): 18-24.)

[9] 白海燕，朱礼军. 关联数据的自动关联构建研究[J]. 现代图书情报技术，2010(2)：44-49. (Bai Haiyan, Zhu Lijun. Research on Automatic Interlinking of Linked Data [J]. New Technology of Library and Information Service, 2010(2): 44-49.)

[10] 马费成，赵红斌，万燕玲，等. 基于关联数据的网络信息资源集成[J]. 情报杂志，2011，30(2)：167-170，175. (Ma Feicheng, Zhao Hongbin, Wan Yanling, et al. Integration of Network Information Resource Based on Linked Data [J]. Journal of Intelligence, 2011, 30(2): 167-170, 175.)

[11] 欧石燕，胡珊，张帅. 本体与关联数据驱动的图书馆信息资源语义整合方法及其测评[J]. 图书情报工作，2014，58(2): 5-13. (Ou Shiyan, Hu Shan, Zhang Shuai. An Ontology & Linked Data Driven Semantic Integrantion Method of Library Information Resources and Its Evaluation [J]. Library and Information Service, 2014, 58(2): 5-13.)

[12] Schmachtenberg M, Bizer C, Paulheim H. Adoption of the Linked Data Best Practices in Different Topical Domains [C]. In: Proceedings of the 13th International Semantic Web Conference, Riva del Garda, Italy. Springer International Publishing, 2014: 245-260.

[13] Auer S, Demter J, Martin M, et al. Lodstats-An Extensible Framework for High-Performance Dataset Analytics [C]. In: Proceedings of the 18th International Conference on Knowledge Engineering and Knowledge Management, Galway, Ireland. Springer Berlin Heidelberg, 2012: 353-362.

[14] Campinas S, Ceccarelli D, Delbru R, et al. The Sindice-2011 Dataset for Entity-Oriented Search in the Web of Data [C]. In: Proceedings of the 1st International Workshop on Entity-oriented Search (EOS), Beijing, China. 2011: 26-32.
[15] Bizer C, Eckert, K, Meusel R, et al. Deployment of RDFa, Microdata, and Microformats on the Web-A Quantitative Analysis [C]. In: Proceedings of the 12th International Semantic Web Conference, Sydney, Australia. 2013: 17-32.
[16] 汪小帆, 李翔, 陈关荣. 网络科学导论[M]. 北京: 高等教育出版社, 2012: 108-115. (Wang Xiaofan, Li Xiang, Chen Guanrong. Network Science: An Introduction [M]. Beijing: High Education Press, 2012: 108-115.)
[17] 汪小帆, 李翔, 陈关荣. 复杂网络: 理论及其应用[M]. 第4版. 北京: 清华大学出版社, 2006: 22-34. (Wang Xiaofan, Li Xiang, Chen Guanrong. Complex Networks: Theory and Its Application [M]. The 4th Edition. Beijing: Tsinghua University Press, 2006: 22-34.)
[18] About the Datahub [EB/OL]. [2015-04-28]. https://datahub.io/about.
[19] Describing Linked Datasets with the Void Vocabulary [EB/OL]. [2015-04-28]. http://www.w3.org/TR/void/.
[20] Heath T, Bizer C. Linked Data: Evolving the Web into a Global Data Space [M]. San Rafael: Morgan & Claypool Publishers, 2011:64.
[21] RKB Explorer [EB/OL]. [2015-04-28]. http://www.rkbexplorer.com/explorer.

作者贡献声明:

谭荧: 设计研究方案, 实验, 撰写论文;
夏立新: 提出研究思路, 修订论文。

收稿日期: 2015-07-20
收修改稿日期: 2015-10-13

Analysis and Visualization of the LOD Network Structure

Xia Lixin　Tan Ying
(School of Information Management, Central China Normal University, Wuhan 430079, China)

Abstract: [**Objective**] This paper aims to analyze the structural features of Linked Open Data (LOD), and the results can be used to guide the organization of linked data in practice. [**Methods**] Describing LOD network with degree distribution, average path length, clustering coefficient and other indexes, this paper compares scale-free network and small-world network in the complex network theory. [**Results**] The structure of LOD network shows a power-law distribution, approximate the scale-free network. The Publication subnet of LOD shows a relatively homogeneous exponential distribution. Two networks both have a short average path length and high clustering coefficient. [**Limitations**] Lack of assigning key nodes to more weight. [**Conclusions**] Small-world phenomenon of LOD can optimize the retrieval efficiency, and scale-free feature will reduce the stability of the entire network.

Keywords: Linked Open Data　Complex network　Network structure　Visualization